

NTRFINDER: A SOFTWARE TOOL TO FIND NESTED TANDEM REPEATS

A. A. MATROUD, M. D. HENDY, AND C. P. TUFFLEY

ABSTRACT. We introduce the software tool NTRFinder to find the complex repetitive structure in DNA we call a nested tandem repeat (NTR). An NTR is a recurrence of two or more distinct tandem motifs interspersed with each other. We propose that nested tandem repeats can be used as phylogenetic and population markers.

We have tested our algorithm on both real and simulated data, and present some real nested tandem repeats of interest. We discuss how the NTR found in the ribosomal DNA of taro (*Colocasia esculenta*) may assist in determining the cultivation prehistory of this ancient staple food crop.

NTRFinder can be downloaded from <http://www.maths.otago.ac.nz/~aamatroud/>.

1. INTRODUCTION

Genomic DNA has long been known to contain *tandem repeats*: repetitive structures in which many approximate copies of a common segment (the *motif*) appear consecutively. Several studies have proposed different mechanisms for the occurrence of tandem repeats [Weitzmann *et al.*, 1997, Wells, 1996], but their biological role is not well understood.

Recently we have observed a more complex repetitive structure in the ribosomal DNA of *Colocasia esculenta* (taro), consisting of multiple approximate copies of two distinct motifs interspersed with one another. We call such structures *nested tandem repeats* (NTRs), and the problem of finding them in sequence data is the focus of this paper. Our motivation is their potential use for studying populations: for example, a preliminary analysis suggests that changes in the NTR in taro have been occurring on a 1,000 year time scale, so a greater understanding of this NTR offers the potential to date the early agriculture of this ancient staple food crop.

The problem of locating tandem repeats is well known, as their implication for neurological disorders [Macdonald *et al.*, 1993, Fu *et al.*, 1992], and their use to infer evolutionary histories has urged some researchers to develop tools to find them. This has resulted in a number of software tools, each of which has its own strengths and limitations. However, the existing tools were not designed to find NTRs, and consequently do not generally find them. In this paper, we present a new software tool, NTRFinder, which is designed to find these more complex repetitive structures.

We report here the algorithm on which NTRFinder is based and report some of the NTRs it has identified, including an even more complex structure where copies of four distinct motifs are interspersed.

2. BACKGROUND AND DEFINITIONS

2.1. Sequences, edit operations and the edit distance. A DNA sequence is a sequence of symbols from the nucleotide alphabet $\Sigma = \{A, C, G, T\}$. We define a DNA *segment* to be a string of contiguous DNA nucleotides and define a *site* to be a component in a segment. For a DNA segment

$$\mathbf{X} = x_1x_2 \cdots x_n,$$

$x_i \in \Sigma$ is the nucleotide at the i -th site and $|\mathbf{X}| = n$ is the length of \mathbf{X} .

Copying errors happen in DNA due to different external and internal factors. These changes include substitution, insertion, deletion, duplication, and contraction. We refer to these as *edit operations*. By giving each type of edit operation some specific weight, we can in principle find a series of edit operations which transform segment x to segment y , whose sum of weights is minimal. We will refer to this sum as the *edit distance*, and denote it by $d(x, y)$. For the purposes of this paper, the edit operations allowed in calculating the edit distance are single nucleotide substitutions, and single nucleotide insertions or deletions (indels), with each given weight 1.

2.2. Classification of Tandem Repeats. Many classifications of tandem repeat schemas have been introduced in the computational biology literature. We list some which are commonly used:

- **(Exact) Tandem Repeats:** An *exact tandem repeat* (TR) is a sequence comprising two or more contiguous copies $\mathbf{XX} \cdots \mathbf{X}$ of identical segments \mathbf{X} (referred to as the *motif*).
- **k -Approximate Tandem Repeats:** A *k -approximate tandem repeat* (k -TR) is a sequence comprising two or more contiguous copies $\mathbf{X}_1\mathbf{X}_2 \cdots \mathbf{X}_n$ of similar segments, where each individual segment \mathbf{X}_i is edit distance at most k from a template segment \mathbf{X} .

- **Multiple Length Tandem Repeats:** A multiple length tandem repeat is a tandem repeat where each repeat copy is of the form $\mathbf{X}\mathbf{x}^n$, where n is a constant larger than one and $d(\mathbf{X}, \mathbf{x})$ is greater than some threshold value k .

Examples:

- **TR:**
AGG AGG AGG AGG AGG. The motif is AGG.
- **1-TR:**
AGG AGC ATG AGG CGG. The motif is AGG.
- **MLTR:**
GACCTTTGG ACGGT ACGGT ACGGT
GACCTTTGG ACGGT ACGGT ACGGT.
The motifs are ACGGT and GACCTTTGG, with $n = 3$.

2.3. Nested Tandem Repeats. In this section we introduce a more complex repetitive structure, the nested tandem repeat (NTR), also referred to as a *variable length tandem repeat* [Hauth and Joseph, 2002]. Let \mathbf{X} and \mathbf{x} be two segments (typically of different lengths) from the alphabet $\Sigma = \{A, C, G, T\}$, such that $d(\mathbf{X}, \mathbf{x})$ is greater than some threshold value k .

Definition 1. An *exact nested tandem repeat* is a string of the form

$$\mathbf{x}^{s_0} \mathbf{X} \mathbf{x}^{s_1} \mathbf{X} \dots \mathbf{X} \mathbf{x}^{s_n},$$

where $n > 1$, $s_i \geq 1$ for each $0 < i < n$, and $s_j \geq 2$ for some $j \in [0, 1, \dots, n]$. The motif \mathbf{x} is called the *tandem repeat* and the motif \mathbf{X} is the *interspersed repeat*. The concatenations of the tandem repeats \mathbf{x}^{s_i} alone, and of the interspersed motifs \mathbf{X} alone, both form exact tandem repeats.

Example: $\mathbf{x} = \text{ACGGT}$, $\mathbf{X} = \text{GACCTTTGG}$, $n = 7$, $s_0 = 0$, $s_1 = 3$, $s_2 = 5$, $s_3 = 2$, $s_4 = 4$, $s_5 = 1$, $s_6 = s_7 = 2$, so

$$\begin{aligned} \mathbf{x}^0 \prod_{i=1}^7 \mathbf{X} \mathbf{x}^{s_i} &= \text{XxxxXxxxxXxxXxxxxXxXxxXxx} \\ &= \text{GACCTTTGG ACGGT ACGGT ACGGT} \\ &\quad \text{GACCTTTGG ACGGT ACGGT ACGGT ACGGT ACGGT} \\ &\quad \text{GACCTTTGG ACGGT ACGGT} \\ &\quad \text{GACCTTTGG ACGGT ACGGT ACGGT ACGGT} \\ &\quad \text{GACCTTTGG ACGGT} \\ &\quad \text{GACCTTTGG ACGGT ACGGT} \\ &\quad \text{GACCTTTGG ACGGT ACGGT.} \end{aligned}$$

In practice we expect any nested tandem repeats occurring in DNA sequences to be approximate rather than exact. In what follows we will write $\tilde{\mathbf{X}}$ to mean an approximate copy of the motif \mathbf{X} , and $\tilde{\mathbf{x}}^s$ to mean an approximate tandem repeat consisting of s approximate copies of the motif \mathbf{x} .

Definition 2. A (k_1, k_2) -*approximate nested tandem repeat* is a string of the form

$$\tilde{\mathbf{x}}^{s_0} \tilde{\mathbf{X}} \tilde{\mathbf{x}}^{s_1} \tilde{\mathbf{X}} \dots \tilde{\mathbf{X}} \tilde{\mathbf{x}}^{s_n},$$

where n and s_i satisfy the same conditions in Definition 1, and $\tilde{\mathbf{x}}^{s_0} \tilde{\mathbf{x}}^{s_1} \dots \tilde{\mathbf{x}}^{s_n}$ is a k_1 -approximate tandem repeat with motif \mathbf{x} , and $\tilde{\mathbf{X}} \tilde{\mathbf{X}} \dots \tilde{\mathbf{X}}$ is a k_2 -approximate tandem repeat with motif \mathbf{X} .

Examples:

- **NTR:**
AGG AGG CTCAG AGG CTCAG AGG AGG AGG CTCAG.
The motifs are AGG, CTCAG.
- **(1, 2)-NTR:**
AGA AGG CTTCG AGG CTCAG AA AGA AGG CTTCG AGG
CTCAG AAG.
The motifs are AGG, CTCAG.

3. RELATED WORK

Various algorithms have been introduced to find exact tandem repeats. Such algorithms were developed mainly for theoretical purposes, namely, to solve the problem of finding squares in strings [Apostolico and Preparata, 1983, Crochemore, 1981, Kolpakov *et al.*, 2001, Main and Lorentz, 1984, Stoye and Gusfield, 2002]. These algorithms are not easily adapted to finding the approximate tandem repeats that usually occur in DNA.

A number of algorithms, [Delgrange and Rivals, 2004, Landau *et al.*, 2001] consider motifs differing only by substitutions, using the Hamming distance as a measure of similarity. Others, e.g. [Benson, 1999, Hauth and Joseph, 2002, Domanić and Preparata, 2007, Sagot and Myers, 1998, Wexler *et al.*, 2005], have considered insertions and deletions by using the edit distance. Most of these algorithms have two phases, a scanning phase that locates candidate tandem repeats, and an analysis phase that checks the candidate tandem repeats found during the scanning phase.

The only algorithm designed to look for NTRs is that of Hauth and Joseph (2002), which searches for tandem motifs of length at most six nucleotides.

4. THE ALGORITHM

In this section we present the algorithm we have developed to search for nested tandem repeats in a DNA sequence. The algorithm requires several preset parameters. These are: k_1 and k_2 which bound the edit distances from the tandem and interspersed motifs; and the motif length bounds $\min_{t_1}, \max_{t_1}, \min_{t_2}, \max_{t_2}$. Other input parameters are discussed below.

Search phase. Our search is confined to seeking NTRs with motifs of length $l_1 \in [\min_{t_1}, \max_{t_1}]$ and $l_2 \in [\min_{t_2}, \max_{t_2}]$. A (k_1, k_2) -NTR must contain a k_1 -TR, so we begin by scanning the sequence for approximate tandem repeats. Several good algorithms, including those of Benson (1999), Wexler *et al.* (2005) and Domanić and Preparata (2007), have been developed to find k_1 -TRs. We have chosen to adapt the algorithm of Wexler *et al.* (2007), where the sequence is scanned by two windows w_1, w_2 of width w , a distance l_1 apart. Wexler's algorithm uses a similarity parameter q with default value $q = 0.5$, which can be reset by the user. The user may set the k_1, k_2 values, preset with default values

$$k_1 = l_1(1 - p_m) + \sqrt{l_1(1 - p_m)p_m}$$

$$k_2 = l_2(1 - p_m) + \sqrt{l_2(1 - p_m)p_m},$$

following Domanić and Preparata (2007), with matching probability p_m given the default value $p_m = 0.8$.

Once a TR has been found and its full extent determined, the right-most copy of the repeated pattern is taken as the current TR motif x , and further approximate copies of x are sought, displaced from the TR up to a distance of \max_{t_2} nucleotides to the right. If no further approximate copies of x are located, this TR is abandoned, and the TR search continues to the right. If a displaced approximate copy of x is observed, then both x and the interspersed segment X are recorded in a list, as we have found a candidate NTR. Further contiguous copies of x are then sought, with the rightmost copy x replacing the previous template motif.

The steps above are repeated with successive motifs x and interspersed segments copied to the list, until no additional copies of the last recorded motif x are found. This search phase is illustrated in Figure 1.

At this point the algorithm builds consensus patterns for x and X using majority rule. After constructing the two consensus patterns the algorithm moves to the verification phase.

Example: An example will help illustrate the procedure. Suppose that S contains an NTR of the form

$$xX_0xxxX_1xxxxxxX_2xxX_3.$$

The algorithm will scan from the left until it locates the tandem repeat consisting of three copies of x between X_0 and X_1 . It will then start searching for additional non-adjacent copies of x to the right, locating the first copy to the right of X_1 . Having found this it will record the intervening segment X_1 , and then continue the tandem repeat search from this point until the full extent of the tandem repeat between X_1 and X_2 is found.

This procedure is repeated once more, locating the tandem repeat between X_2 and X_3 , recording the segment X_2 , and then searching for further copies to the right. At this point no more copies of x are found, and the process of verification begins. The segments X_0, X_3 and the initial copy of x are found during this stage.

Verification phase: Each candidate NTR is checked to determine whether it meets the NTR definition. This is accomplished by aligning the candidate NTR region, together with a margin on either side of it, against the consensus motifs x and X , using the nested wrap-around dynamic programming algorithm of Matroudi *et al.* (2010). This has complexity $O(n|x||X|)$, where n is the length of the NTR region and $|x|$ and $|X|$ are the length of the tandem motif and the length of the interspersed motif respectively.

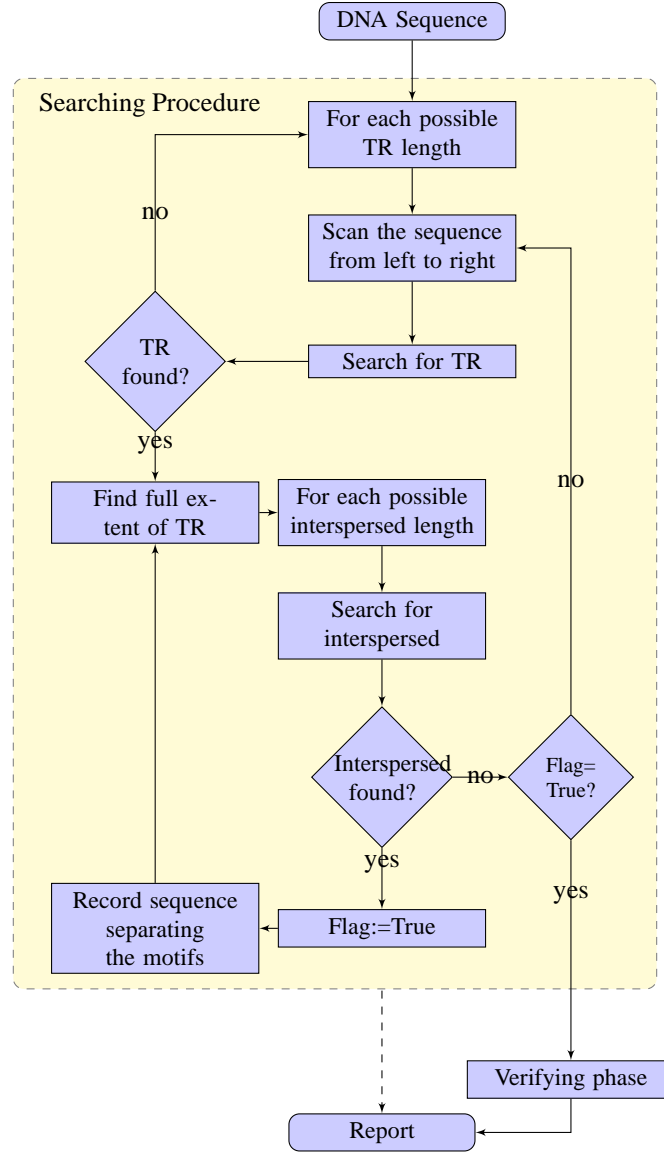


FIGURE 1. Flowchart of the NTRFinder algorithm.

5. RESULTS

5.1. Tests on real sequence data. We have implemented our algorithm and carried out searches for NTRs on some DNA sequences taken from GenBank. The size ranges used for this search were $[\min_{t_1}, \max_{t_1}] = [\min_{t_2}, \max_{t_2}] = [2, 100]$, with the parameters k_1 , k_2 and q left set to their default values. Some NTR regions found by our software are listed in Table 1.

5.2. More complex structures. In addition to the nested tandem repeats in Table 1, NTRFinder also reported an NTR in *Linum usitatissimum* (accession number gi—164684852—gb—EU307117.1—) which on further analysis by hand turned out to have a more complex structure. The IGS region of the rDNA of this species contains an NTR with four motifs interspersed with each other. The four motifs are $w=GTGCGAAAAT$, $x=GCGCGCCAGGG$, $y=GCACCCATAT$, and $z=GCGATTTTG$ and the structure of the NTR is

$$\prod_{i=1}^{25} w^{q_i} x^{r_i} z^{s_i} y^{t_i},$$

where $q_i \in \{1, 2, 3\}$; $r_i \in \{1, 2\}$; $s_i \in \{0, 1\}$; $t_i \in \{0, 1\}$.

5.3. Running time. The running time for NTRFinder searching some sequences from GenBank is shown in Figure 2. It can be seen that the run time is approximately linear in the length of the sequence. However, it must be noted that

Species Accession number	tandem motif x interspersed motif X	$ x $ $ X $	start index end index	# x # X
<i>N. sylvestris</i> X76056.1	$x = \text{AGGACATGGC}$ $X = \text{CATGGCACGAGTC}$	10 13	960 2,111	53 26
<i>B. juncea</i> X73032.1	$x = \text{GGACGTCCCGCGTGACACAGAC}$ $X = \text{CACAGACGGTCGACCTGGACGACCTGCGTG}$	21 30	1,403 2,605	51 7
<i>B. oleracea</i> X60324.1	$x = \text{GGACAGTCCTCGTGGGCGAAAAATCACCCAC}$ $X = \text{GGATAGTCCACGGGAAGGGCCAACGTGCTGATATGCGTACTGAC}$	30 44	1,256 3,341	32 20
<i>B. rapa</i> S78172.1	$x = \text{GGATCAGTACAC}$ $X = \text{GTCCACGGGAAGGGCCAACATGCTGATATGTGTAATACACGGACA}$	12 45	385 1,337	20 8
<i>B. campestris</i> S78172.1	$x = \text{GGACGTCTTTGTGTGTCTGAC}$ $X = \text{GGACACACGGACACACACGGACACGCCAGGGGAAGGGCCAGCGTGTGTCTGAC}$	21 51	1,558 2,580	37 8
<i>C. esculenta</i> Not published	$x = \text{TCGCACAGCCG}$ $X = \text{TTCTGGGCAAAACGGCTGGGTGACGTGTCTGAACTGGCCAGCTGGTTCTG}$	11 48	725 2384	94 12
<i>D. melanogaster</i> AE014296.4	$x = \text{TGCCCCAGT}$ $X = \text{TGCTGCTCGCCTGGC}$	9 15	4,215,779 4,215,899	7 6
<i>H. sapien X chromosome</i> AL672277.21	$x = \text{CT}$ $X = \text{CACAAGGAGCTGCTCTCCTCCTTCTCTGTGTGAGACGTGTGTGTCTGTCTTT}$	2 55	35,471 35,711	360 8
<i>H. sapien X chromosome</i> AL683871.15	$x = \text{GATA}$ $X = \text{TGATGGTAATAGATACATACTTAGTA}$	4 27	111,705 113,805	147 56

TABLE 1. Nested tandem repeats found in some sequences from GenBank and an additional unpublished sequence (*C. esculenta*).

the run time depends not only on the length of the input sequence, but also on the number of tandem and nested tandem repeats found in the sequence. The program spends most of the time verifying any tandem repeats found.

6. DISCUSSION

In the last decade a number of software tools to find tandem repeats have been introduced; however, little work exists on more complex repetitive structures such as nested tandem repeats. The problem of finding nested tandem repeats is addressed in this study. The motivation for our study is the potential use of NTRs as a marker for genetic studies of populations and of species.

We have done some analysis on the nested tandem repeat in the intergenic spacer region in *C. esculenta* (taro), noting some variation in the NTRs derived from domesticated varieties sourced from New Zealand, Australia and Japan. Further varieties are currently being analysed. By considering some edit operations such as deletion, mutation, and duplication we can align the nested tandem repeat regions of each pair of sequences. The alignment score can then be considered as a measure of distance between both sequences. In particular they appear to share some common inferred histories of the development of the NTRs from a simpler structure of two motifs. The edit operations appear to be occurring on a 1,000

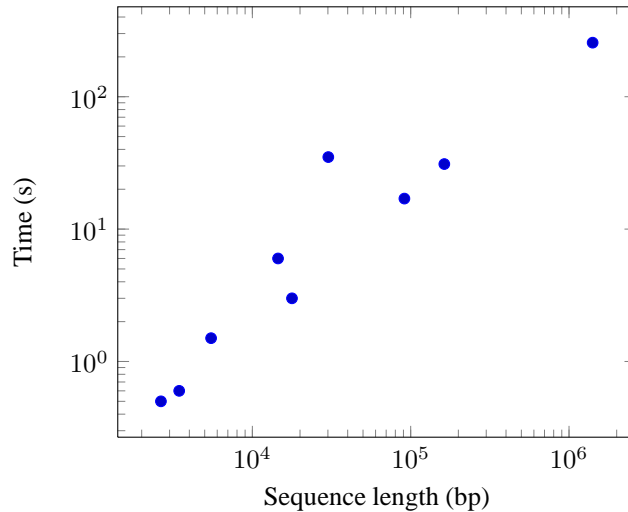


FIGURE 2. Running time of NTRFinder (on a Pentium Dual core T4300 2.1 GHz) plotted against segment length on a log-log scale. The search was performed on segments of different lengths, with the minimum and maximum tandem repeat lengths set to 8 and 50 respectively. The distribution suggests the running time is approximately linear with sequence length.

year timescale, so this analysis offers the potential to date the prehistory of the early agriculture of this ancient staple food crop.

7. CONCLUSION

The nested tandem repeat structure is a complex structure that requires further analysis and study. The number of copy variants in the NTR region and the relationships between these copies might suggest a tandem repeat generation mechanism. In this paper, we have introduced a new algorithm to find nested tandem repeats. The first phase of the algorithm has $O(n(\max_{t_1})(\max_{t_2}))$ time complexity, while the second phase (the alignment) needs $O(n(\max_{t_1})(\max_{t_2}))$ space and time, where n is the length of the NTR region, and \max_{t_1} , \max_{t_2} are the maximum allowed lengths of the tandem and interspersed motifs.

ACKNOWLEDGEMENTS

Andrew Clarke and Peter Matthews, for providing data and useful background about Taro, and Hussain Matawa, for assisting in the development of the program interface.

Funding: This project was partially funded by the Allan Wilson Centre for Molecular Ecology and Evolution.

REFERENCES

- [Apostolico and Preparata, 1983] Apostolico, A., Preparata, F. P., (1983) Optimal Off-Line Detection of Repetitions in a String, *Theor. Comput. Sci.*, **22**, 297-315.
- [Benson, 1999] Benson G., (1999) Tandem repeats finder: a program to analyze DNA sequences, *Nucl. Acids Res.*, **27**, 2, 573-580.
- [Boeva *et al.*, 2006] Boeva, V. and Regnier, M. and Papatsenko, D., Makeev, V., (2006) Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression, *Bioinformatics*, **22**, 6, 676-684.
- [Buard and Jeffreys, 1997] Buard, J. and Jeffreys, A.J., (1997) Big, bad minisatellites, *Nat Genet*, **15**, 4, 327-328.
- [Crochemore, 1981] Crochemore M., (1981) An Optimal Algorithm for Computing the Repetitions in a Word, *Inf. Process. Lett.*, **12**, 5, 244-250.
- [Delgrange and Rivals, 2004] Delgrange, O., Rivals, E., (2004) STAR: an algorithm to Search for Tandem Approximate Repeats, *Journal of Comp. Bio.*, **20**, 16, 2812-2820.
- [Domanić and Preparata, 2007] Domanić, N. O., Preparata, F. P., (2007) A Novel Approach to the Detection of Genomic Approximate Tandem Repeats in the Levenshtein Metric, *Journal of Comp. Bio.*, **14**, 7, 873-891.
- [Fu *et al.*, 1992] Fu, YH. *et al.*, (1992) An unstable triplet repeat in a gene related to myotonic muscular dystrophy, *Science.*, **255**, 5049, 1256-1258.
- [Hauth and Joseph, 2002] Hauth, A. M., Joseph, D., (2002) Beyond tandem repeats: complex pattern structures and distant regions of similarity, *ISMB.*, 31-37.
- [Kolpakov *et al.*, 2001] Kolpakov, R., Kucherov, G., Logiciel, T. G., (2001) Finding approximate repetitions under Hamming distance, *Theor. Comput. Sci.*, **22**, 6, 170-181.
- [Landau *et al.*, 2001] Landau, G. M. and Schmidt, J. P. and Sokol, D., (2001) An Algorithm for Approximate Tandem Repeats, *Journal of Comp. Bio.*, **8**, 1, 1-18.
- [Macdonald *et al.*, 1993] Macdonald, M. E. *et al.*, (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group, *Cell.*, **72**, 971-983.
- [Main and Lorentz, 1984] Main, M. G., Lorentz, R. J., (1984) An $O(n \log n)$ Algorithm for Finding All Repetitions in a String, *J. Algorithms*, **5**, 3, 422-432.
- [Matroud *et al.*, 2010] Matroud, A. A., Hendy, M. D., Tuffley, C. P., (2010) An Algorithm to Solve the Motif Alignment Problem for Approximate Nested Tandem Repeats, *RECOMB-CG*, 188-197.
- [Sagot and Myers, 1998] Sagot, M. F., Myers, E. W., (1998) Identifying Satellites and Periodic Repetitions in Biological Sequences, *Journal of Comp. Bio.*, **5**, 3, 539-554.
- [Stoye and Gusfield, 2002] Stoye, J., Gusfield, D. (2002) Simple and flexible detection of contiguous repeats using a suffix tree, *Theor. Comput. Sci.*, **270**, 1-2, 843-856.
- [Verkerk *et al.*, 1991] Verkerk, A. *et al.*, (1991) Friedreich's Ataxia: Autosomal Recessive Disease Caused by an Intronic GAA Triplet Repeat Expansion, *Science.*, **65**, 5, 905-914.
- [Weitzmann *et al.*, 1997] Weitzmann, M. N. and Woodford, K. J. and Usdin, K. (1997) DNA Secondary Structures and the Evolution of Hypervariable Tandem Arrays, *J. Biol. Chem.*, **272**, 14, 9517-9523.
- [Wells, 1996] Wells, R. A. (1996) Molecular Basis of Genetic Instability of Triplet Repeats, *J. Biol. Chem.*, **271**, 6, 2875-2878.
- [Wexler *et al.*, 2005] Wexler, Y. and Yakhini, Z. and Kashi, Y., Geiger, D., (2005) Finding Approximate Tandem Repeats in Genomic Sequences, *Journal of Comp. Bio.*, **12**, 7, 928-942.

ALLAN WILSON CENTRE FOR MOLECULAR ECOLOGY AND EVOLUTION, MASSEY UNIVERSITY, PRIVATE BAG 11 222, PALMERSTON NORTH 4442, NEW ZEALAND

INSTITUTE OF FUNDAMENTAL SCIENCES, MASSEY UNIVERSITY, PRIVATE BAG 11 222, PALMERSTON NORTH 4442, NEW ZEALAND

INSTITUTE OF FUNDAMENTAL SCIENCES, MASSEY UNIVERSITY, PRIVATE BAG 11 222, PALMERSTON NORTH 4442, NEW ZEALAND